EVALUATING SECONDARY DATA SOURCES FOR RANDOM DIGIT DIALING SAMPLES

J. Michael Brick, David Judkins, Jill Montaquila, David Morganstein, and Gary Shapiro, Westat
J. Michael Brick, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Sample design, stratification, rare population, and listed telephone numbers

1. Introduction

In recent years, more and more information on individual telephone numbers and on telephone exchanges can be obtained from publicly available databases. This paper evaluates strategies for improving the efficiency of random digit dialing (RDD) household surveys by using these data in a stratified design.

The approach is to stratify either telephone exchanges (the 3 numbers following the area code) or individual numbers depending on the nature of the intelligence. A sample design assigns a sampling rate to each stratum based upon a number of parameters. The following parameters determine the optimum sampling rates for a stratified RDD design: percent of phone numbers in the stratum; percent of the stratum that are members of the study population; and, the ratio of costs of screening and locating within the stratum to the costs to conduct the interview. When a stratum of numbers or exchanges can be identified that contains a significant percentage of a rare population and the percentage of the rare population is high within the stratum, then the use of oversampling can be considered. The optimal degree of oversampling depends upon the proportion of the study population that belongs to the stratum and on the relative costs of screening for residential status, screening for the specific study population and interviewing.

The data sources we examine fall into two categories. One type is data available at an aggregate level. These files contain information derived from models that use the latest (currently 1990) census data. The models take statistics known at an aggregate level, such as a census tract, and project them in two ways. For one, they attempt to adjust for changes which have occurred since the latest Census. Second, they project from a somewhat geographically compact unit, such as a tract, to a telephone exchange. Data items of this type include race, ethnicity and poverty status, all items initially known from the Census at an aggregate level. Obviously, this modeling results in estimates which may differ substantially from the current characteristics of households in a telephone exchange. Further, since they are aggregate estimates, there will still be variation in survey eligibility within strata.

The second type of data is given at the telephone number level. These data do not involve the kind of modeling required by the first type. These data come from secondary sources and provide information about the telephone number or about the household and its members. Variables of this type include age of household members, vehicle registration and listed status of the telephone number. The quality issues that affect their usefulness in the sample design are completeness (the percent of households for which information is known) and accuracy (the extent to which the information is correct). These data items may be out of date or may exist for only a small portion of the rare population.

In Section 2, we present the statistical theory that combines the sampling error computation with a cost model to arrive at an optimal sampling rate. Section 3 lists the data sources that we examined in this study. Section 4 reviews the distributions of these data. In Section 5, we review the accuracy of the various data items. In Section 6, we discuss our recommended strategy for each rare population, given various survey costs. Finally, Section 7 provides a summary.

2. Theory of Oversampling High Concentration Strata

This section discusses the theory underlying oversampling from strata believed to have high concentrations of a rare population (for a more complete treatment see Mohadjer, 1988 and Mohadjer and West, 1992). We describe an approach that establishes sampling rates that minimize the sampling errors of the estimates for a fixed budget. We describe how the sampling errors are affected by a disproportionate stratified sample.

It is also necessary to understand the cost structure of a household RDD survey, with screening and extended interview expenses, to arrive at the answer. With any RDD survey, considerable effort is needed to call and screen phone numbers to determine those that are residential. When sampling for a rare population, a screening interview is attempted with all residential contacts to determine if the phone number has identified a household that either is or contains a member of that rare population. The ratio of all the costs needed to locate a rare population member to the costs to conduct the extended interview plays an important role in determining the optimal sampling rate.

The survey might be designed to make estimates only for the rare population or it might also need to make estimates for the entire population while making cost-effective and efficient estimates for the rare subpopulation. Although the appropriate strategy for each objective has many common features, we discuss each separately to capture their unique requirements.

Estimates Needed Solely for Rare Population

A common misconception is that it is always efficient to increase the sampling rate in a stratum if a large percentage of the population in that stratum are members of the rare population. For example, if identifiable parts of a city are known to be almost entirely from a desired minority race or ethnic group, then it may be tempting to take a significant portion of (if not the entire) sample from those areas. While this concentration is a necessary condition to support oversampling, it is not sufficient. A second condition is that a significant portion of the entire rare population comes from that stratum. See Waksberg (1973) and Kalton et. al (1986).

We now discuss a cost model useful for assessing the most efficient extent of oversampling. The choice of a correct cost model can be difficult. The presented model should be regarded as an example of what might be used, but it is not necessarily the best model for any particular situation. The analyses we present in this paper do not all use this model.

Consider a stratification scheme with strata of varying concentrations of one or more rare populations. Our model of the total cost, C, of an RDD survey that is stratified into two strata and has a fixed number of completed extended interviews, n, is:

$$C - c_0 = \sum_h n_h \{(t_h - 1) r_h c_d + r_h c_s + c_e\}, \qquad (1)$$

where

n_h is the number of completed extended interviews in stratum h, with $n = n_1 + n_2$;

t_h is the average number of telephone numbers dialed to obtain one completed household screener interview;

r_h is the average number of households with completed screeners required to obtain one completed extended interview in stratum h;

c_0 is the fixed cost of conducting the survey (planning, sampling training, etc.);

c_d is the average cost of data collection for telephone numbers that do not result in screening interviews;

c_s is the average cost of one completed residential screening interview; and

c_e the average cost of one extended interview.

The parameters in the model need further explanation. The average cost of data collection for telephone numbers that do not result in completed screener interviews, c_d, includes distinct costs. It can be represented as a convex combination of two costs,

$$c_d = \alpha c_n + (1 - \alpha) c_u, \qquad (2)$$

where

c_n is the average cost of eliminating one nonresidential telephone number;

c_u is the average cost of unproductive attempts to obtain a completed screener interview with a nonresponding residential telephone number; and

α is the number of nonresidential telephone numbers (for which calls are made), divided by the sum of the number of nonresidential telephone numbers and the number of nonresponding residential telephone numbers.

The cost of eliminating a nonresidential telephone number, c_n, includes the tritone purging and the cost of interviewers dialing into business and nonworking telephone numbers that are not identified in the tritone purge efforts. The cost is averaged over all nonresidential telephone numbers. The average cost of unproductive attempts to obtain a completed screener, c_u, includes unsuccessful attempts to convince households to respond to the screening interview and unsuccessful attempts to determine residential status (mainly repeated ring/no answer and answering machine results). Here the cost is averaged over all nonresponding residential numbers. Clearly, the data collection protocol could have a significant effect on these costs.

The fixed cost, c_0, is assumed to be minor in the development that follows. The rationale is that once this method of sampling is established there is little additional cost associated with applying it rather than not applying it. This is true of organizations that do RDD samples regularly, but may not be true of those organization that rarely do this work.

The other two parameters worth describing in more detail are the multipliers that account for completion rates. The first is the average number of telephone numbers needed to get one household that completes a screener, t_h, in stratum h. We can write this number as $t_h = (v_h \gamma_h)^{-1}$, where v_h is the residency rate in stratum h and γ_h is screener response rate in stratum h.

The second parameter is the average number of households with completed screeners required to get one completed extended interview in stratum h, r_h. We can write this number as $r_h = (\beta_h \xi_h)^{-1}$ where β_h is the eligibility rate (the proportion of households with eligible members), and ξ_h is the extended interview completion rate. For example, a survey may only be interested in sampling veterans of the US military and β_h would be the estimated proportion of households in stratum h that have veterans.

Applying standard sampling theory for optimal allocation (e.g., Cochran, 1977) for the case of two strata, when the stratum variances are equal, the optimal ratio of the sampling fraction in the first stratum to the second stratum for the number of completed extended interviews is

$$\lambda = \left\{ \frac{(t_2-1)r_2(\alpha c_n + (1-\alpha)c_u) + r_2 c_s + c_e}{(t_1-1)r_1(\alpha c_n + (1-\alpha)c_u) + r_1 c_s + c_e} \right\}^{\frac{1}{2}}. \quad (3)$$

The expression is a familiar result of optimal allocation, however, it is in terms of the number of completed extended interviews not the number of sampled telephone numbers. In other words, it is the ratio of $\frac{n_1}{N_1}$ to $\frac{n_2}{N_2}$, where N_1 is the total number of persons in the rare population group in stratum 1 and N_2 is the number of persons in the rare population group in stratum 2. However, the ratio of the sampling rates between the two strata for the sampling of telephone numbers is also λ, $\left(\text{i.e.,} \lambda = \frac{m_1}{M_1}\left(\frac{m_2}{M_2}\right)^{-1} \right)$ because $N_h = \frac{M_h}{t_h r_h}$ and $n_h = \frac{m_h}{t_h r_h}$, where M_h is the estimated number of telephone numbers in stratum h.

The reduction in the variance of the estimates using the optimal allocation given by (3) relative to the variance obtained under an allocation proportional to the number of telephone numbers in the stratum[1], holding total costs fixed, can be evaluated. For allocation proportional to the number of telephone numbers, the total cost can be written as

$$C^* = n_2^{pr}\left[\left(\frac{M_1 t_2 r_2}{M_2 t_1 r_1}\right)\{(t_1-1)r_1 c_d + r_1 c_s + c_e + (t_2-1)r_2 c_d + r_2 c_s + c_e\}\right], \quad (4)$$

where the superscript "*pr*" indicates allocation proportional to the number of telephone numbers. Using the fixed cost, C^*, the optimal allocation in the second stratum is given by

$$n_2^{opt} = C^*\left[\left(\frac{\lambda M_1 t_2 r_2}{M_2 t_1 r_1}\right)\{(t_1-1)r_1 c_d + r_1 c_s + c_e + (t_2-1)r_2 c_d + r_2 c_s + c_e\}\right]^{-1}. \quad (5)$$

Using expressions (4) and (5), the ratio of the variance under optimal allocation to the variance under proportional allocation is given by

$$\phi = \frac{n_2^{pr}}{n_2^{opt}} \frac{\frac{M_1}{\lambda t_1 r_1} + \frac{M_2}{t_2 r_2}}{\frac{M_1}{t_1 r_1} + \frac{M_2}{t_2 r_2}}, \quad (6)$$

In Section 3 below, we evaluate alternative strata definitions with varying degrees of concentration for the various rare populations under study.

Estimates Needed for Both Rare and Entire Population

Consider the situation when estimates are required for the entire population, as well as, for a rare subgroup of the population. Inevitably the screening needed to locate the rare population will locate more non-rare population members than necessary. The differential sampling rates used in the two strata induce a design effect which is accepted in the estimation of the rare population but which is inefficient (and unnecessary) for estimates of the non-rare population. Accordingly, non-rare population members identified in the high concentration stratum should be subsampled so that all non-rare population members are sampled at the same rate, regardless of the stratum in which they are located.

[1] This design is the basis for comparison because it is the average design that results when stratification by listed status is not used.

3. Evaluation for Various Rare Populations

In this section we discuss the sources of information for use in defining strata for various populations. In general, we desire to stratify telephone numbers into the following strata:

- Those with a large expected concentration of a rare population;
- Those with a small expected concentration of a rare population; and
- Those for which no information is known.

3.1 Micro Level Data

Listed Phone Numbers

In most RDD telephone surveys, at least those conducted for the federal government, telephone numbers in at least partially listed 100 banks are sampled with equal probability. An option offered by some vendors of telephone numbers is to select banks of 100 numbers with differential probabilities using the number of listed telephone numbers in the bank as a measure of size. Casady and Lepkowski (1993) discuss this approach, but it has about the same efficiency (considering both cost and variance) as equal probability sampling.

Another approach is to stratify telephone numbers by whether or not they are listed and independently sample at possibly different rates in the two strata (first suggested by Judkins, 1996). This approach has some obvious advantages. Listed telephone numbers are much more likely to be residential so the cost of finding a residence is much lower in the stratum of listed numbers. In addition, households with listed telephone numbers are more likely to cooperate with most surveys, especially if methods such as mailing advance letters to those households are used to boost the response rates.

Youth (Ages 16-24)

We designed a survey targeted at male youth, aged 16-24 who either had graduated high school or were still in high school, but had not dropped out of high school. Based on statistics from the Current Population Survey (CPS), about 22 percent of households contain youths aged 16 to 24, of whom about 11 percent have dropped out of high school. We did not compute the percent of these that contained males in the targeted age range, but it was probably about 70 percent, so the overall rarity of the targeted population was about 14 percent or 1 in 7. Given a desire for a sample size of 16,400, if the residential rate is 51 percent and the response rate is 75 percent, then the overall raw sample size for the RDD sample must be about 306,000 phone numbers. This is obviously a very expensive proposition. Given recent advances in the Genesys database owned by MSG, we decided to test whether the Genesys system which provides an indicator of household member ages was accurate enough for use in an oversampling design.

3.2 Aggregate Data

Race/Ethnicity

Often, estimates by race/ethnicity are of interest to analysts of data from RDD household surveys. One example can be taken from the National Household Education Survey (NHES) conducted by Westat for the National Center for Education Statistics (NCES). For this survey, estimates of the care and educational experiences of children are frequently produced by race/ethnicity, with separate categories for non-Hispanic blacks and for Hispanics. In order to attain adequate precision for these estimates, it is necessary to oversample blacks and Hispanics. However, accurate race/ethnicity data are not available at the telephone number level. Therefore, the mechanism for oversampling blacks and Hispanics that has been used for the NHES has been to oversample telephone exchanges predicted to have high concentrations of blacks and Hispanics, where the concentration is determined using Decennial Census data to model the expected current concentration for the geographic area.

4. Distributions of Data on Telephone Numbers an Exchanges

In this section we discuss the distributions of data about telephone numbers and exchanges. In a survey conducted in 1999, about 55 percent of all the cooperating households had listed telephone numbers even though only 33 percent of all eligible telephone numbers (those in banks with at least one listed telephone number) are listed.

4.1 Micro Level Data Used for Sampling Male Youth

Four strata were established in the Genesys database for the purpose of oversampling males aged 16 to 24 who have not dropped out of high school. These strata are shown in Table 1 along with the prevalence of each stratum within the standard list-assisted universe for RDD sampling (i.e., phone numbers in 100 banks with at least one listed residential phone number). Even if the age and sex information were not very good, one would hope that the combination of strata 2 through 4 would offer some efficiency since at least all the phone numbers are listed residential phone numbers. A large portion of the cost in an RDD survey is screening out the nonworking and nonresidential phone numbers from stratum 1.

4.2 Aggregate Data for Sampling Blacks and Hispanics

In this section we discuss the concentration of rare populations estimated for the telephone exchange.

The MSG sampling frame contains estimates modeled from the most recent available Decennial Census (currently, the 1990 Census) of the race/ethnicity distributions of persons in the telephone exchange. For the purpose of oversampling blacks and Hispanics, stratification schemes based on the minority concentration of the exchange were considered. Each stratification scheme involved the creation of a "high minority" stratum and a "low minority" stratum. Five different definitions were considered for NHES for the high minority stratum: At least 10 percent black or at least 10 percent Hispanic; at least 20 percent black or at least 20 percent Hispanic; at least 30 percent black or at least 30 percent Hispanic; at least 20 percent black or Hispanic; and at least 30 percent black or Hispanic.

Table 2 gives the percentages of the total population, of telephone numbers, and of listed telephone numbers in the high minority stratum for each of the stratification schemes considered.

5. Accuracy of the Data

In this section we assess the accuracy of information reported for exchanges or phone numbers by comparing the frame data used in stratification to survey data provided by the household for sampling blacks and Hispanics. Table 3 compares expected proportions based on models to those actually achieved in the NHES, for one set of definitions of high minority/low minority strata. The correspondence is quite good. For example, the model indicated that 25.6 percent of the households in the high minority stratum would be black. The proportion of households in the NHES sample that were black was slightly higher, 28.8 percent. These figures provide strong evidence that the model estimates are quite accurate, and that estimated gains due to stratification and oversampling are approximately correct.

6. Recommended Oversampling Rates

In this section we provide advice about oversampling rates as a function of the cost ratio of screening to interviewing.

Earlier we pointed out that in multipurpose surveys compromises are needed to satisfy conflicting requirements. For example, the optimal allocation for estimating characteristics of blacks is somewhat different than for Hispanics and very different than for overall totals. If precise estimates are desired for both groups and for the total, a compromise allocation is needed. The final allocation can often be reached by estimating effective sample sizes for each required domain and total using the formula given in Section 3.

6.1 Listed

In our evaluation for the NHES, we found that for overall estimates (i.e., estimates of all race/ethnicity combined), the optimum is to sample listed numbers at about 1.2 times the rate of unlisted numbers. The gain in efficiency, however, is less than 2 percent.

6.2 Male Youth

To evaluate the utility of list-based oversampling for this domain, we asked MSG to merge onto a special sample of phone numbers the list characteristics available at the time the sample had been drawn. This special sample consisted of residential phone numbers that were found through standard list-assisted RDD sampling for a prior survey where the residents answered our screening questions about the age and sex of the residents. Among those who were eligible for the prior survey, only those who actually completed the extended interview were included in this analysis. This special sample contained a total of 88,332 residential phone numbers that were complete or ineligible for the prior survey. Since we have directly talked with the residents of the homes owning these 88,332 phone numbers, we can provide a very accurate assessment of the completeness and accuracy of the Genesys list for sampling males aged 16 to 24 who have not dropped out of high school.

Table 4 shows how the completes and ineligibles from the prior survey would have been stratified by MSG if we had chosen to stratify the sample. Values of r_h ranged from about 2 to 30. These are the screening ratios among cooperative residential phone numbers. In the stratum where MSG believed there were eligible males, 1 in 2.3 households did indeed report an eligible male. This was a far better screening ratio than the rate of 1 in 30.2 households in the listed stratum where MSG believed there were no eligible males. From that perspective, the list clearly has some predictive power. Also, the screening ratio in the unlisted households and listed households with missing age data were similar to each other and very different from the ratio for listed housing with age data.

However, only 14.5 percent of the eligible males were in stratum 4. Because of this very low coverage of the targeted population and the cost structure for the survey, there was very little benefit to be gained from optimal allocation. Taking the universe of cooperative residential phone numbers as the base, a telephone interview with an eligible male would cost about as much as the cost to screen enough raw phone numbers to find 2.28 cooperative residential phone numbers when there was no stratification. Applying the formula from Section 2, the optimal relative sampling rate for each stratum was calculated, and is shown in Table 4 where this was defined to be $(n_h/N_h)/(n_0/N_0)$. Relative to the unlisted stratum, it is optimal to slightly undersample listed phone numbers with missing age data, more strongly undersample listed phone numbers with contra-indicative age data, and to oversample listed phone numbers with eligible age data by a factor slightly under 2.

If the cost of stratification is small enough to be ignored, then the use of the oversampling rate is projected to reduce the total cost of data collection by 5 to 6 percent. This may be a small understatement of the savings because we restricted the matching sample to completes and ineligibles from the prior survey and thus could not quantify the cost savings from reduced dialing of nonresidential phone numbers. As noticed above, oversampling listed numbers can improve survey efficiency by 1 or 2 percent, so the total efficiency gain might be as much as 7 percent.

6.3 Race/Ethnicity

We used data from the NHES:1999 to examine the effectiveness of oversampling based on the minority concentration in the telephone exchange. Table 5 gives the estimated distributions of the population across minority strata for the alternative in which the high minority stratum is defined as the set of telephone exchanges having at least 20 percent black or at least 20 percent Hispanic. For these strata, Table 5 also gives the estimated average amount of screening required to sample a person by race/ethnicity domain. As shown in Table 5, large proportions of the black and Hispanic populations are in the high minority stratum under each scheme.

Table 6 gives the optimal oversampling ratios for the NHES for each race/ethnicity subgroup, for each of the stratification schemes considered in Section 4.2. The most extreme case is for sampling blacks using the stratification scheme in which a telephone exchange is defined as high minority if at least 10 percent of the population is black or at least 10 percent is Hispanic. In this case, it is optimal to sample telephone numbers in the high minority stratum by a factor of about 2.8, and to substantially undersample telephone numbers in the low minority stratum, provided the aim is to screen for blacks only. This results in a sampling rate for the high minority stratum about 20 times as high as for the low minority stratum. With the same definition of strata, the optimal allocation for sampling persons in the "other" race/ethnicity subgroup calls for under sampling telephone numbers in the high minority stratum (an oversampling rate of 0.8), which results in a sampling rate for telephone numbers in the low minority stratum about double that for those in the high minority stratum. Obviously for a survey such as the NHES for which estimates are required for both the rare populations and for the total population, there would need to be subsampling of the other households in high minority strata so that final sample of other households would be nearly equi-probability sample.

Table 6 also gives the expected design effects due to oversampling, if the specified oversampling ratios are used. The design effects in the table are for the optimal oversampling ratio for the given type of allocation. For example, design effects of 1.00 are shown for "overall allocation". However, if the optimal Black allocation were used (2.8 oversampling ratio for stratification scheme 1), then the design effect for overall statistics would be higher. For sampling blacks using stratification scheme 1, although the optimal allocation calls for sampling telephone numbers in the high minority stratum at about 20 times the rate of telephone numbers in the low minority stratum, this variation in rates for sampling telephone numbers is expected to increase the variance of Black statistics by only about 10 percent compared to a simple random sample of the same size. This is due to the fact that a large proportion of the black population is in the high minority stratum. We pursued a strategy for NHES in which we defined strata by listed/not listed in addition to high minority/low minority. Therefore, we did not estimate variance gains that would be achieved by applying the oversampling rates in the table.

7. Conclusions

Most of the new data on telephones and exchanges led to only modest improvements in variance in most situations. However, techniques to use these data are easy to implement and may be worthwhile in many situations.

When using micro level data for a subgroup like males aged 16-24, variance gains are slight because only a small proportion of the young males are in listed households identified as containing young males. Nonetheless, since the implementation costs are also low this strategy may be useful for future RDD surveys of youth.

Oversampling telephone numbers with listed addresses results in modest improvements in efficiency. This methodology can be usefully applied for most RDD surveys.

Using aggregate data for race/ethnicity results is beneficial, especially if the survey is only interested in a minority group such as blacks or Hispanics. Current research suggests that combining data on listed status and aggregate data on race/ethnicity will achieve the greatest variance gains.

8. References

Casady, R.J., and Lepkowski, J.M. (1993). Stratified Telephone Survey Design. *Survey Methodology*, **19**(1) p. 103-113.

Cochran, W. (1977). *Sampling Techniques*. John Wiley & Sons.

Judkins, D. (1996). Directions for Research on the National Immunization Survey. Final report. Research Triangle Institute. Research Triangle Park, NC.

Kalton, G. and Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society*, Series, A, 149, **1**, pp. 65-82.

Mohadjer, L. (1988). Stratification of Prefix Areas for Sampling Rare Populations. *Telephone Survey Methodology*. Chapter 10, Wiley, pp. 161-174.

Mohadjer, L. and West, J. (1992). Effectiveness of Oversampling Blacks and Hispanics in the NHES Field Test: National Household Education Survey Technical Report. Washington, D.C.: U.S. Department of Education.

Waksberg, J. (1973). The Effects of Stratification With Differential Sampling Rates on Attributes on Subsets of the Population. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 429-434.

Table 1. Genesys sampling strata for oversampling male youth

Stratum	Description	Percentage of cooperative residential RDD phone numbers
1	Unlisted	43.4
2	Listed but either age or sex data missing on all occupants	24.4
3	Listed with age and sex data – Genesys indicates no males aged 16 to 24	30.0
4	Listed with age and sex data – Genesys indicates males aged 16 to 24	2.2

Table 2. Percentages of the targeted rare populations, the total population, of telephone numbers, and of listed telephone numbers in the high minority stratum under alternative stratification schemes

Definition of high minority stratum	Percent of black population in high minority stratum	Percent of Hispanic population in high minority stratum	Percent of population in high minority stratum	Percent of telephone numbers in high minority stratum	Percent of listed telephone numbers in high minority stratum
Stratification scheme 1: At least 10 percent black or at least 10 percent Hispanic	89.5	87.4	51.4	52.8	48.2
Stratification scheme 2: At least 20 percent black or at least 20 percent Hispanic	73.6	71.7	33.0	33.7	29.3
Stratification scheme 3: At least 30 percent black or at least 30 percent Hispanic	57.2	56.2	21.2	21.7	17.8
Stratification scheme 4: At least 20 percent black or Hispanic	81.2	79.7	40.1	41.5	36.6
Stratification scheme 5: At least 30 percent black or Hispanic	66.3	66.5	27.3	28.2	23.6

SOURCE: MSG 1st quarter 2000 database.

Table 3. Expected and observed percent, by strata and race/ethnicity for scheme 2

	Black		Hispanic		Other	
Stratum	Proportion observed on the frame	Proportion observed at screening	Proportion observed on the frame	Proportion observed at screening	Proportion observed on the frame	Proportion observed at screening
Overall	11.5	12.5	10.8	11.4	77.7	76.1
High minority*	25.6	28.8	23.5	24.1	50.9	47.1
Low minority	4.5	4.3	4.6	5.0	90.9	90.7

*High minority is defined as "At least 20 percent black or at least 20 percent Hispanic."

SOURCE: U.S. Department of Education, National Center for Education Statistics (NCES), National Household Education Survey (NHES), 1999 and tabulations of the Genesys Sampling systems, Marketing systems Group (MSG) 1st Quarter 2000 database.

Table 4. Distribution of males aged 16 to 25 who are not high school dropouts across Genesys sampling strata

Stratum	Percent of eligible males ages 16 to 24	Screened households to get one interview (r_h)	Optimal relative sampling rate
Unlisted	47.4%	13.7	1.00
Listed with missing age or sex data	23.3%	15.7	0.94
Listed with age and sex data, no males 16 to 24	14.9%	30.2	0.70
Listed with males 16 to 24	14.5%	2.3	1.87
	100.0%		

Table 5. Estimated percentages of the population in each stratum, overall and by race/ethnicity, and average amount of screening required to sample a person in the race/ethnicity domain

	Percent of total	Black, non-Hispanic		Hispanic		Other	
Stratum	population in stratum	(At percent)	r_h**	(At percent)	r_h**	(At percent)	r_h**
Overall	100.0	100.0	8.0	100.0	8.8	100.0	1.3
Minority stratum							
High minority*	33.4	77.0	3.5	70.8	4.1	10.7	2.1
Low minority	66.6	23.0	23.3	29.2	20.0	79.3	1.1

*High minority is defined as "At least 20 percent black or at least 20 percent Hispanic."

**The values of r_h given here do not reflect subsampling of households for sampling of persons for extended interviews or the extended interview completion rate.

SOURCE: U.S. Department of Education, National Center for Education Statistics (NCES), National Household Education Survey (NHES), 1999.

Table 6. Optimal oversampling ratios and expected design effects due to oversampling under the optimal allocation for each race/ethnicity subgroup, by stratification scheme

Stratification scheme	Optimal oversampling ratio				Expected design effect due to oversampling			
	Black allocation	Hispanic allocation	Other allocation	Overall allocation	Black allocation	Hispanic allocation	Other allocation	Overall allocation
1: At least 10 percent black or at least 10 percent Hispanic	2.8	2.5	0.8	1.0	1.10	1.09	1.01	1.00
2: At least 20 percent black or at least 20 percent Hispanic	2.3	2.2	0.8	1.0	1.13	1.12	1.01	1.00
3: At least 30 percent black or at least 30 percent Hispanic	2.2	2.2	0.7	1.0	1.14	1.13	1.01	1.00
4: At least 20 percent black or Hispanic	2.4	2.3	0.8	1.0	1.12	1.12	1.01	1.00
5: At least 30 percent black or Hispanic	2.2	2.2	0.7	1.0	1.14	1.14	1.02	1.00